Rendering Realistic Subject-Dependent Expression Images by Learning 3DMM Deformation Coefficients

Claudio Ferrari(✉), Stefano Berretti, Pietro Pala, and Alberto Del Bimbo

Media Integration and Communication Center,
University of Florence, Florence, Italy
claudio.ferrari@unifi.it

Abstract. Automatic analysis of facial expressions is now attracting an increasing interest, thanks to the many potential applications it can enable. However, collecting images with labeled expression for large sets of images or videos is a quite complicated operation that, in most of the cases, requires substantial human intervention. In this paper, we propose a solution that, starting from a neutral image of a subject, is capable of producing a realistic expressive face image of the same subject. This is possible thanks to the use of a particular 3D morphable model (3DMM) that can effectively and efficiently fit to 2D images, and then deform itself under the action of deformation parameters learned expression-by-expression in a subject-independent manner. Ultimately, the application of such deformation parameters to the neutral model of a subject allows the rendering of realistic expressive images of the subject. Experiments demonstrate that such deformation parameters can be learned from a small set of training data using simple statistical tools; despite this simplicity, very realistic subject-dependent expression renderings can be obtained. Furthermore, robustness to cross dataset tests is also evidenced.

Keywords: 3D morphable model · Deformation components learning · Facial expression synthesis

1 Introduction

In Computer Vision there is an increasing interest in developing methods for either *recognizing* or *synthesizing* expressions in an automatic way. In fact, this has both theoretical interest in disciplines as different as Cognitive Sciences, Medicine or Psychology, as well as in practical applications, like surveillance by analysis of human emotional state, monitoring for fatigue detection, gaming or Human Computer Interaction, to cite a few. While for long time automatic analysis of facial expressions from images and videos has been based on the design of hand-crafted features, now the success of neural networks, and deep learning

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11130, pp. 441–455, 2019.
https://doi.org/10.1007/978-3-030-11012-3_34

solutions in particular, has drastically changed the scenario: the idea is to let the network learn the low- and intermediate-level features that are best suited to describe the training data, and then use them in any classification or recognition task. This moves most of the criticisms to the networks design and the collection of the data used for their training. In doing so, the amount of the data and their variability play a fundamental role in learning significant representations. In the case of facial expressions, this has some additional difficulties since obtaining large quantities of ground truth data with accurate expression labels is a complicated and time consuming task if executed by human annotators. Thus, an idea that is making its way is to synthetically generate such training data. To this end, solutions based on parametric models, like the 3D Morphable Model (3DMM) [5] and its variants are among the most promising. The idea here is to fit such model to 2D target images so as to reconstruct a coarse 3D shape of the face. Then, this 3D face model can be deformed to exhibit a target expression and render a corresponding image. Of course, this process requires the deformation components that change the neutral model to an expressive one are known for each expression. This, by itself, is not an easy task since most of the 3DMMs have been trained without using any expressive scan [7]. Some recent works also applied Generative Adversarial Networks (GANs) for the task of generating expressive face images from neutral ones [19,29]. However, also in this case, 3DMMs can play a role for generating the images used for GANs training.

Rendering expressive images of a subject starting from his/her neutral one using parametric face models has potential applications also in designing advanced interfaces and serious games [22]. For example, a desktop system could use an avatar to interact with the user adapting the avatar's expression to that of the user; similarly, two avatars could be used in a virtual call simulating the expression of the interacting people. A training scenario appears also realistic, where disabled people or people recovering after a disease or injuries that compromised their facial mimic (*e.g.*, a stroke) could use a virtual assistant to learn reproducing facial expressions in a correct way [3]. People affected by autism syndrome could also benefit from an application that helps them in reproducing expressions. This could be done by starting from a model representing the neutral face of the subject and then by producing different expressions on it [37].

In this paper, we develop on the idea of automatically synthesize images of expressive faces. To this end, we start with a particular variant of the 3DMM, which is characterized by its capability of reproducing facial expressions starting from the average model. This is possible thanks to two specific aspects of this 3DMM (called DL-3DMM [15]): *(i)* differently from most existing 3DMMs it is trained also with 3D expressive facial scans; *(ii)* its deformation components are learned as a dictionary of atoms using a *dictionary learning* approach; differently from the standard approach that learns deformation components by Principal Component Analysis (PCA) so that each component acts globally on the model, the atoms identified by the dictionary learning solution capture quite well local deformations of the face. This 3DMM can be efficiently fit to a target 2D face

image using a closed form solution generating a coarse 3D model of the target subject. Our goal here is to deform such 3D neutral model so as to realize a given expression of the subject ultimately rendering a 2D expressive image. To this end, we design a learning procedure that identifies the weights of the atoms corresponding to each prototypical expression. The procedure is composed of two main steps: first, the 3DMM is fit to a face image in neutral expression, producing a person-specific 3D reconstruction; then, this reconstruction is used to fit an expressive face image of the same subject and the deformation parameters are collected. This allows us to separate between the deformations that model identity traits and the ones modeling expressions. Once all these parameters are collected, we look for recurrent patterns among them that identify prototypical expressions and use such parameters to control the 3DMM deformation and generate expressive models. Such parameters are expression-specific, but can also be mixed together so as to generate more complex expressions. Experiments show that this strategy permits us to recover such parameters pretty easily, and that we can effectively generate expressive and realistic models also in a cross-dataset fashion. In particular, the main contributions of this work are as follows:

- We propose a simple yet effective framework that enables the extrapolation of 3DMM parameters that control expression-specific deformations, and successfully apply them to generate expressive renderings starting from face images in neutral expression;
- We showcase the potential and versatility of the DL-3DMM in handling and generating expressions;
- We demonstrate the generalization capability of our solution by showing that more complex expressions can be generated by combining different prototypical expression parameters.

The rest of the paper is organized as follows: In Sect. 2, the works in the literature that are most closely related to our proposed solution are discussed; In Sect. 3, we summarize the 3DMM used in this work and the characteristics that make it effective in modeling facial expressions; In Sect. 4, we present the methods used to learn the deformation coefficients related to each expression; These coefficients are then used to generate expressions starting from the neutral 3DMM for new identities; a qualitative evaluation is reported in Sect. 5; Finally, discussion and conclusions are drawn in Sect. 6.

2 Related Work

In the following, first, we report on the solutions that define and use a 3DMM to derive the 3D face model of a target subject starting from his/her 2D neutral image; then, we summarize some methods that learn modes of deformations to transform a neutral 3D model to an expressive one.

Blanz and Vetter [5] first presented a complete solution to derive a 3DMM by transforming the shape and texture from a training set of 3D face scans into a vector space representation based on PCA. However, the training dataset had

limited face variability (200 neutral scans of young Caucasians), thus reducing the capability of the model to generalize to different ethnicity and non-neutral expressions. Despite these limitations, the 3DMM has proved its effectiveness in image face analysis, inspiring most of the subsequent work. The 3DMM was further refined into the Basel Face Model by Paysan et al. [28]. This offered higher shape and texture accuracy thanks to a better scanning device, and a lower number of correspondence artifacts using an improved registration algorithm based on the non-rigid Iterative Closest Point (ICP) [2]. However, since non-rigid ICP cannot handle large missing regions and topological variations, expressions were not accounted for in the training data also in this case. In addition, both the optical flow used in [5] and the non-rigid ICP method used in [1,28] were applied by transferring the vertex index from a reference model to all the scans, so that the choice of the reference face can affect the quality of the detected correspondences, and the resulting 3DMM. The work by Booth et al. [8] introduced a pipeline for 3DMM construction. Initially, dense correspondence was estimated applying the non-rigid ICP to a template model. Then, the so called LSFM-3DMM was constructed using PCA to derive the deformation basis on a dataset of 9,663 scans with a wide variety of age, gender, and ethnicity. Though the LSFM-3DMM was built from the largest dataset compared to the current state-of-the-art, the face shapes were still in neutral expression.

Following a different approach, Patel and Smith [27] showed that Thin-Plate Splines (TPS) and Procrustes analysis can be used to construct a 3DMM. In [12], Cosker et al. described a framework for building a dynamic 3DMM, which extended static 3DMM construction by incorporating dynamic data. This was obtained by proposing an approach based on Active Appearance Model and TPS for non-rigid 3D mesh registration and correspondence. Results showed this method overcomes optical flow based solutions that are prone to temporal drift. Brunton et al. [9], instead, proposed a statistical model for 3D human faces in varying expression. The approach decomposed the face using a wavelet transform, and learned many localized, decorrelated multilinear models on the resulting coefficients. In [24], Lüthi et al. presented a Gaussian Process Morphable Model (GPMM), which generalizes PCA-based Statistical Shape Models (SSM).

3DMM has been used at coarse level for face recognition and synthesis. In one of the first examples, Blanz and Vetter [6] used their 3DMM to simulate the process of image formation in 3D space, and estimated 3D shape and texture of faces from single images for face recognition. Later, Romdhani and Vetter [31] used the 3DMM for face recognition by enhancing the deformation algorithm with the inclusion of various image features. In [35], Yi et al. used the 3DMM to estimate the pose of a face image with a fast fitting algorithm. This idea was extended further by Zhu et al. [38], who proposed fitting a dense 3DMM to an image via Convolutional Neural Network. Grupp et al. [16] fitted the 3DMM based exclusively on facial landmarks, corrected the pose of the face and transformed it back to a frontal 2D representation for face recognition. Hu et al. [17] proposed a Unified-3DMM that captures intra personal variations due to illumination and occlusions, and showed its performance in 3D-assisted 2D face

recognition for scenarios where the input image is subjected to degradations or exhibits intra-personal variations. Recent solutions also used deep neural networks to learn complex non-linear regressor functions mapping a 2D facial image to the optimal 3DMM parameters [14,33].

In all these cases, the 3DMM was used mainly to compensate for the pose of the face, with some examples that performed also illumination normalization. Expressions were typically not considered. Indeed, the difficulty in making 3DMM work properly in fine face analysis applications is confirmed by the existence of very few methods that use 3DMM for expression recognition [4,10]. Among the few examples, Ramanathan et al. [30] constructed a 3D Morphable Expression Model incorporating emotion-dependent face variations in terms of morphing parameters that were used for recognizing four emotions. Ujir and Spann [34] combined the 3DMM with Modular PCA and Facial Animation Parameters (FAP) for facial expression recognition, but the model deformation was due more to the action of FAP than to the learned components. In [13], Cosker et al. used a dynamic 3DMM [11] to explore the effect of linear and non-linear facial movement on expression recognition through a test where users evaluated animated frames. Huber et al. [20] proposed a cascaded-regressor based face tracking and a 3DMM shape fitting for fully automatic real-time semi dense 3D face reconstruction from monocular in-the-wild videos. The Dictionary Learning based 3DMM (DL-3DMM) proposed by Ferrari et al. [15] was one of the most promising in producing realistic facial expressions from the mean model. This is possible thanks to a dense alignment procedure based on landmarks, face partitioning and resampling, which allows expressive scans are enrolled in the training. This 3DMM has been used to enhance facial expression and action unit recognition from 2D images and videos with state-of-the-art performance on benchmark datasets.

3 3D Morphable Model

From the discussion of existing solutions for generating a 3DMM, it is quite evident the presence of some aspects that play a major relevance in characterizing the different solutions: (1) the human face variability captured by the model, which directly depends on the number and heterogeneity of training examples; (2) the capability of the model to account for facial expressions; also this feature of the model directly derives from the presence of expressive scans in the training. One of the few 3DMM existing in the literature that exposes both these features is the Dictionary Learning based 3DMM (DL-3DMM) proposed by Ferrari et al. [15]. Since our contribution mainly develops on this model, to make the paper as self-contained as possible, below we describe the peculiar features that make this particular 3DMM formulation suitable for our purposes.

3.1 DL-3DMM Construction

The first problem to be solved in the construction of a 3DMM is the selection of an appropriate set of training data. This should include sufficient

variability in terms of ethnicity, gender, age, so as to enable the model to include a large variance in the data. Apart for this, the most difficult aspect in preparing the training data is the need to provide dense, *i.e.*, vertex-by-vertex, alignment between the 3D scans. In the original work of Blanz and Vetter [5] this was solved with the optical-flow method that provided reasonable results just in the case of neutral scans of the face. Several subsequent works used non-rigid variants of the Iterative Closest Point (ICP) algorithm, thus solving some problem related to the optical-flow, but without the explicit capability of addressing large facial expressions in the training data. The dense alignment of the training data for the DL-3DMM was obtained with a different solution based on the detection of landmarks of the face, and their use for partitioning the face into a set of non-overlapping regions, each one identifying the same part of the face across all the scans. Re-sampling the internal of the region based on its contour, a dense correspondence is derived region-by-region and so for all the face. Such method showed to be robust also to large expression variations as those occurring in the Binghamton University 3D facial Expression (BU-3DFE) database [36]. This latter dataset was used in the construction of the DL-3DMM.

Once a dense correspondence is established across the training data, these are used to estimate a set of deformation components that will be used to generate novel shapes. In the classic 3DMM framework [5], new 3D shapes $\mathbf{S}$ are generated by deforming an average model $\mathbf{m}$ with a linear combination of a set of M principal components $\mathbf{C}$, usually derived by PCA as follows;

$$\mathbf{S} = \mathbf{m} + \sum_{i=1}^{|M|} \mathbf{C}_i \alpha_i. \tag{1}$$

The DL-3DMM is instead constructed by learning a dictionary of deformation components exploiting the *Online Dictionary Learning for Sparse Coding* technique [25]. Learning is performed in an unsupervised way, without exploiting any knowledge about the data (*e.g.*, identity or expression labels). Then, the average model is deformed using the dictionary atoms $\mathbf{D}_i$ in place of $\mathbf{C}_i$ in Eq. (1).

Dictionary learning is usually cast as an ℓ_1-regularized least squares problem [25]; however, the sparsity induced by the ℓ_1 penalty to the dictionary atoms, can lead to discontinuous components and ultimately in a noisy or punctured 3D shape. To address this issue, the dictionary learning is formulated as an *Elastic-Net* regression, mitigating the sparsity effect of the ℓ_1 penalty with an ℓ_2 regularization that forces smoothness. By defining $\ell_{1,2}(\mathbf{w}_i) = \lambda_1 \|\mathbf{w}_i\|_1 + \lambda_2 \|\mathbf{w}_i\|_2$, where λ_1 and λ_2 are, respectively, the sparsity and regularization parameters, the problem can be formulated as (using N training scans):

$$\min_{\mathbf{w}_i,\, \mathbf{D}} \frac{1}{N} \sum_{i=1}^{N} \Big(\|\mathbf{v}_i - \mathbf{D}\mathbf{w}_i\|_2^2 + \ell_{1,2}(\mathbf{w}_i) \Big), \tag{2}$$

where $\mathbf{v}_i \in \mathbb{R}^{3m}$ is the vector of deviations between scan i and the average model (being m the number of points in the scans), the columns of the dictionary

$\mathbf{D} \in \mathbb{R}^{3m \times k}$ are the basis components, $\mathbf{w}_i \in \mathbb{R}^k$ are the coefficients of the dictionary learning, and k is the number of basis components of the dictionary. Note that the number of components (dictionary atoms) is fixed and pre-determined. The coefficients vector $\mathbf{w} \in \mathbb{R}^k$ provides an estimate of the degree of importance that each atom had in reconstructing the training set; in comparison with the classic framework based on PCA, these can be interpreted similarly to the eigenvalues. A favorable characteristic of the DL-3DMM is that, oppositely to PCA, larger dictionaries lead to more accurate reconstructions and are likely to include sparser and complementary atoms; this facilitates the identification of the atoms that involve particular face areas. More details on the dictionary learning procedure can be found in [15].

The average model $\mathbf{m}$, the dictionary $\mathbf{D}$ and $\mathbf{w}$, constitute the DL-3DMM.

3.2 DL-3DMM Fitting

Fitting a 3DMM to a 2D face image allows a coarse 3D reconstruction of the face. To this end, estimating the 3D pose of the face, and the correspondence between 3D and 2D landmarks are prerequisites. In order to estimate the pose, a set of 49 facial landmarks $\mathbf{l} \in \mathbb{R}^{2 \times 49}$ is detected on the 2D face image using the technique proposed in [21], while an equivalent set of vertices $\mathbf{L} \in \mathbb{R}^{3 \times 49}$ is manually annotated on the average 3D model. Under an affine camera model [26], the relation between $\mathbf{L}$ and $\mathbf{l}$ is:

$$\mathbf{l} = \mathbf{A} \cdot \mathbf{L} + \mathbf{T}. \tag{3}$$

The affine matrix is directly estimated with a closed-form least squares solution since, by construction, facial landmark detectors do not permit outliers or unreasonable arrangement of the landmarks. The 2D translation is instead recovered as $\mathbf{T} = \mathbf{l} - \mathbf{A} \cdot \mathbf{L}$. The estimated pose $\mathbf{P}$ is represented as $[\mathbf{A}, \mathbf{T}]$ and used to map each vertex of the 3DMM onto the image.

Using the learned dictionary $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_k]$, the average model is non-rigidly transformed such that the projection minimizes the error in correspondence to the landmarks. The coding is formulated as the solution of a regularized *Ridge-Regression* problem:

$$\arg\min_{\boldsymbol{\alpha}} \left\| \mathbf{l} - \mathbf{PL} - \sum_{i=1}^{k} \mathbf{P}\mathbf{d}_i(\mathbf{I}_v)\alpha_i \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \circ \mathbf{w}^{-1} \right\|_2, \tag{4}$$

where $\circ$ is the Hadamard product and $\mathbf{I}_v$ are the indices that correspond to the vertices of the landmarks in the 3D model. By defining $\mathbf{X} = \mathbf{l} - \mathbf{L}$ and $\mathbf{Y} = \mathbf{PD}(\mathbf{I}_v)$, the solution can be found in closed form as follows:

$$\boldsymbol{\alpha} = \left(\mathbf{Y}^T\mathbf{Y} + \lambda \cdot \mathrm{diag}(\mathbf{w}^{-1}) \right)^{-1} \mathbf{Y}^T\mathbf{X}. \tag{5}$$

Again, for a detailed description of the procedure the reader can refer to [15]. A fitting example obtained using this solution is shown in Fig. 1; the 3D model is deformed according to the target face image, the vertices of the model can be projected onto the face image exploiting the estimated pose so that we can sample its texture.

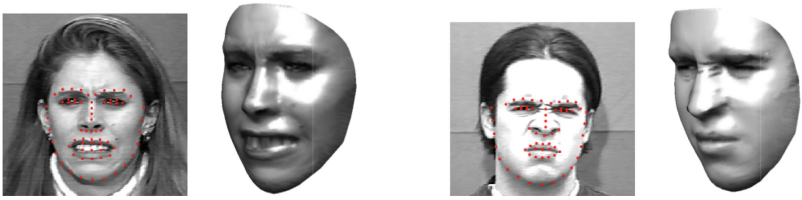

Fig. 1. Examples of DL-3DMM fitting on expressive face images from the Cohn-Kanade (CK+) dataset [23]

4 Learning Expression Coefficients

Given the DL-3DMM as described above, the result of the fitting procedure is a set of coefficients $\boldsymbol{\alpha}$ that are used to deform the average model using Eq. (1). Considering a generic face image, the latter coefficients codify the global shape deformation (*i.e.*, the identity) along with other deformations (*i.e.*, expressions). Our main goal is to derive the set of coefficients that reproduce expressions; in order to do so, we first need to isolate the identity component from the deformation. To this aim, we first fit the DL-3DMM to a face image in neutral expression to account for the identity and obtain the coefficients $\boldsymbol{\alpha}_{id}$; subsequently, the fitted model is used in place of the average model to fit an expressive face image of the same subject. In this way, we obtain a set of coefficients $\boldsymbol{\alpha}_{expr}$ that codify the expression. The procedure is depicted in Fig. 2. The final and crucial step is to find a recurrent pattern in the $\boldsymbol{\alpha}_{expr}$ coefficients, separately for each expression. To this end, we propose to investigate and compare the appropriateness of different methods using: *(i)* statistical indicators; *(ii)* regressors.

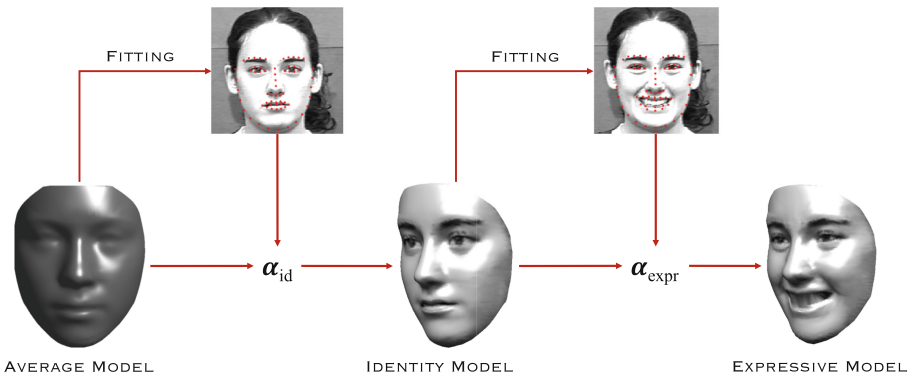

Fig. 2. Workflow of the proposed procedure to extract the expression-specific deformation coefficients from the DL-3DMM fitting

Statistical Indicators. First, we have investigated some basic statistical indicators, namely *mean*, *median* and *mode*. Best results have been obtained with the latter, which we estimated using the *mean-shift* algorithm. Using a Gaussian kernel

$$K(x_i - x) = e^{-c||x_i - x||^2},$$

a centroid x_i at iteration t is updated with $x_i^{t+1} = x_i^t + m(x)$, at iteration $t+1$, with

$$m(x_i) = \frac{\sum_{x_j \in N(x_i)} K(x_j - x_i) x_j}{\sum_{x_j \in N(x_i)} K(x_j - x_i)}.$$

In this latter equation, $N(x_i)$ is a neighborhood of x_i, that is the set of points such that $K(x_i) \neq 0$, and $m(x)$ is the *mean shift vector*. The centroid updating is repeated till the convergence of $m(x)$. The only parameter used in this algorithm is the *bandwidth*, *i.e*, the radius of the gaussian region. In our case, we search for the centroid best representing the data distribution. To this end, we started with a fixed *bandwidth* and repeated the mean shift iteration by increasing the radius; the procedure terminates when an individual point is returned; this centroid is assumed as the vector representing the data distribution of a given expression.

At this stage, we also used the mean-shift algorithm to investigate the data distribution. To this end, first, we fixed the *bandwidth*, applied the algorithm, took the resulting number of centroids and counted how many samples fall in the region of influence of the centroids. Table 1 reports the results for each expression. It can be observed as the first centroid, located at the maximum peak of the data distribution, is the most representative of the samples: arguably, this is due to the fact the other maxima capture possible outliers or errors included in the dataset.

Table 1. For each expression, the number of centroids found by fixing the *bandwidth*, and the percentage of vectors that fall in the region of the first centroid are reported. Expressions in the Cohn-Kanade (CK+) dataset have been used here

Expression	#Centroids	% vectors first centroid
Angry	4	93%
Contempt	5	73%
Disgust	2	98%
Fear	2	96%
Happy	5	92%
Sadness	3	93%
Surprise	2	95%

As a further analysis, we iterated the algorithm by augmenting the *bandwidth* so as to find two centroids. Then, we have compared the faces obtained by applying both the weight vectors corresponding to the two maxima for each expression. Results indicated that the deformed faces obtained from the weight vector of the first maximum are the same as those obtained using a single maximum; applying the weight vector corresponding to the second maximum, instead, resulted in non-realistic faces.

Regressors. In the following, we model the problem of estimating the deformation coefficients of the 3DMM as a regression one, using the Support Vector Regression (SVR) technique. A cross validation process has been used to determine the train/test splits. The coefficients vectors have been used as "multi-labels" that are predicted using a *multi-output* regressor, which repeats the estimate for each component of the array. The regressor is controlled by parameters that do not depend on the dimensionality of the feature space. In our case, a 4-fold cross validation has been performed to determine the best kernel and the values of the parameters C and ϵ of the regressor.

As a result, for both the methods, we obtained for each expression a set of coefficients $\boldsymbol{\alpha}_{est}$ that allow the application of an expression to a subject-specific model in neutral expression.

5 Experimental Results

In this Section, we first describes the dataset adopted for the experimental evaluation (Sect. 5.1), then we provide a qualitative evaluation (Sect. 5.2), of the different modalities we have used for model parameter estimation.

5.1 Dataset

The experiments have been performed on the Extended Cohn-Kanade (CK+) dataset [23], which includes about 600 sequences of 123 subjects showing 7 different expressions, namely *Disgust, Surprise, Angry, Sadness, Fear, Contempt, Happy*; for each sequence the neutral (first) and expressive (last) frames have been used. The DL-3DMM has been deformed to each of these frames using the fitting procedure illustrated in Sect. 3.2; we used a dictionary of 300 atoms. A subset of the 123 individuals has been used to learn the parameters so as to test on different identities. For neutral frames, these coefficients capture the shape information of the individual; for expressive scans, we first deformed the 3DMM on the neutral frame of the same subject, then from this to the expressive frame, following the procedure of Fig. 2. In this way, the coefficients capture the shape deformation to pass from a neutral to an expressive scan for a specific identity.

5.2 Qualitative Results

In order to derive qualitative results, we fitted the DL-3DMM to some neutral faces of the dataset and applied the estimated deformation coefficients $\boldsymbol{\alpha}_{est}$ so as to generate expressive scans for each expression. Figure 3 shows some examples of generated expressive renderings starting from the neutral one and applying the deformation. The magnitude of the deformations can be controlled with a parameter λ, useful to emphasize subtle expressions that do not sufficiently change the neutral face (*e.g.*, contempt). The expressive models generated from the neutral 3DMM according to the learned deformation vectors are rendered for qualitative evaluation. Some examples can be appreciated in Fig. 3; starting

Fig. 3. Qualitative examples of synthetically generated expressive renderings of two subjects from the CK+ dataset. The leftmost column reports the 3DMM fitting to the neutral face images of the two subjects, while the other columns report the models derived from the neutral ones by applying the deformation coefficients corresponding to each expression, from *disgust* to *happy*, left-to-right

from the neutral expression, we can effectively generate expressive renderings applying the expression-specific parameters separately.

Figure 4 shows another interesting application of our method, that is the generation of complex expressions by combining the parameters of the single prototypical expressions. This feature allows us to mix an arbitrary number of expressions and further demonstrates the meaningfulness of the estimated parameters. The examples in Fig. 4 are generated using a combination of 2 (top row) or 3 (bottom row) basic expressions. A drawback of this application is that if the weights of the single expressions are not balanced, the final model can result noisy or excessively deformed.

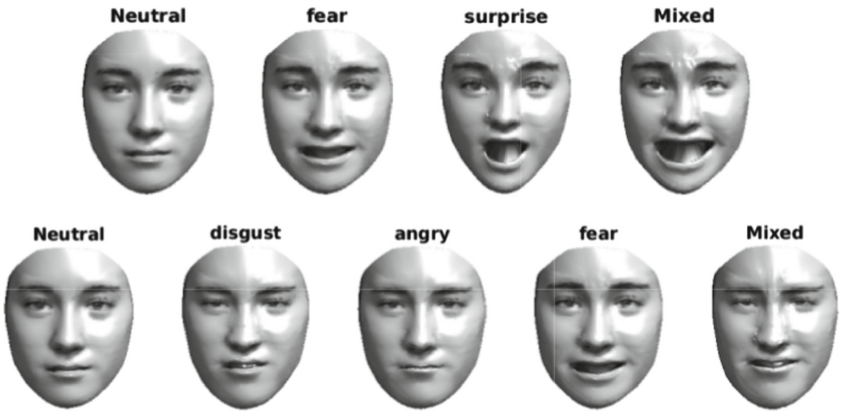

Fig. 4. Qualitative examples of mixed deformations on one subject of the CK+ dataset; two (top row) or three (bottom row) prototypical expressions have been used

In Fig. 5, we show a comparison between the different techniques used to estimate the $\boldsymbol{\alpha}_{est}$ coefficients; the generated images are rather similar to each other,

even using basic statistical indicators as the mean. This suggests us that the elements of the dictionary are effective in separating the identity and expression components and that our methodology allows us to easily extrapolate expression-specific patterns within the deformation coefficients.

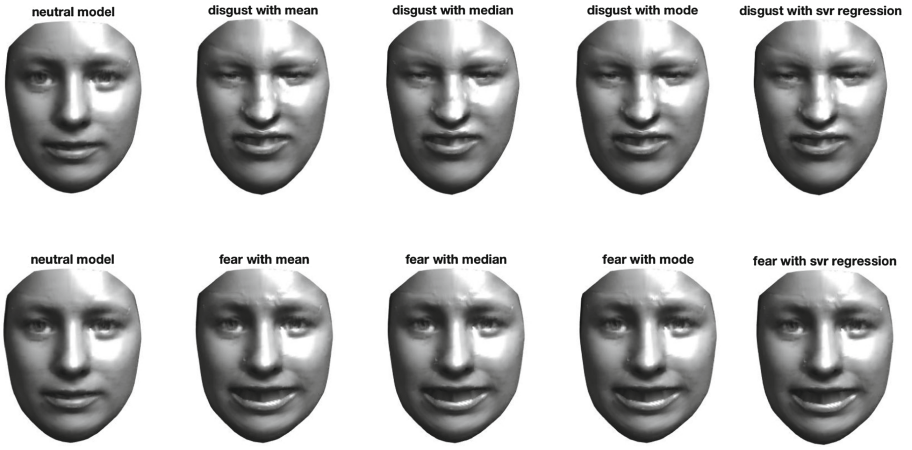

Fig. 5. Qualitative comparison of the different parameters estimation methods

Finally, Fig. 6 shows the application of our expression transfer method to face images coming from different datasets, demonstrating the generalization capability of our approach in a cross-dataset scenario. Specifically, Fig. 6 (top row) shows a face image from the Bosphorus dataset [32], while in Fig. 6 (bottom row) a face image coming from the Labeled Faces in The Wild dataset (LFW) [18] is shown. The former is a 3D face analysis dataset and comprises face images along with their 3D models captured in controlled conditions; the LFW, instead, is composed of "in the wild" face images and is used to address the face verification problem. For both the examples, we are able to transfer the expression of the subject from neutral to any of the learned expressions; this because the 3DMM is independent from the dataset which is applied to.

Fig. 6. Cross-dataset evaluation of the proposed method on sample images from the Bosphorus (top row) and LFW (bottom row)

6 Conclusions

In this paper, we have proposed a method to isolate the expression-specific deformation parameters of a 3DMM and applied them to synthetically generate expressive renderings of subjects in neutral expression. We exploited a peculiar 3DMM implementation based on a dictionary learning technique, able to reproduce expressions thanks to the inclusion of expressive scans in the training set. We showed that our two-step 3DMM fitting methodology is effective in removing the identity component from the 3DMM fitting, and that expression-specific recurrent patterns can be easily found within the parameters used to fit the subject-specific model to its own expressive image. Moreover, the recovered parameters can be effectively mixed so as to generate more complex expressions. However, our solution is not exempt from limitations: first, expressions might be more or less subtle; this means that they must be weighted accordingly in order not to produce exaggerated or imperceptible deformations. Another issue arose is that the textured renderings might result somewhat unnatural at times when trying to generate expressions that are very diverse from the neutral one. Indeed, we can assume that even a very slight expressiveness might be present in "neutral" frames. As a future work, we are considering an extension of the technique to the texture component of the images.

Acknowledgments. The authors would like to thank Gabriele Barlacchi, Francesco Lombardi, Alessandro Sestini, and Alessandro Soci for developing and experimenting part of the code used in this work.

References

1. Amberg, B., Knothe, R., Vetter, T.: Expression invariant 3D face recognition with a morphable model. In: IEEE International Conference on Automatic Face and Gesture Recognition (2008)
2. Amberg, B., Romdhani, S., Vetter, T.: Optimal step nonrigid ICP algorithms for surface registration. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1–8. Minneapolis, MN, June 2007
3. Baranyi, R., Willinger, R., Lederer, N., Grechenig, T., Schramm, W.: Chances for serious games in rehabilitation of stroke patients on the example of utilizing the Wii Fit Balance Board. In: IEEE International Conference on Serious Games and Applications for Health (SeGAH), pp. 1–7, May 2013. https://doi.org/10.1109/SeGAH.2013.6665319
4. Bejaoui, H., Ghazouani, H., Barhoumi, W.: Fully automated facial expression recognition using 3D morphable model and mesh-local binary pattern. In: Blanc-Talon, J., Penne, R., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2017. LNCS, vol. 10617, pp. 39–50. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70353-4_4
5. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: ACM Conference on Computer Graphics and Interactive Techniques (1999)
6. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Trans Pattern Anal. Mach. Intell. **25**(9), 1063–1074 (2003)

7. Booth, J., Antonakos, E., Ploumpis, S., Trigeorgis, G., Panagakis, Y., Zafeiriou, S.: 3D face morphable models "in-the-wild". In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5464–5473, July 2017. https://doi.org/10.1109/CVPR.2017.580
8. Booth, J., Roussos, A., Zafeiriou, S., Ponniahand, A., Dunaway, D.: A 3D morphable model learnt from 10,000 faces. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5543–5552 (2016)
9. Brunton, A., Bolkart, T., Wuhrer, S.: Multilinear wavelets: a statistical shape space for human faces. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 297–312. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_20
10. Chang, F.J., Tran, A., Hassner, T., Masi, I., Nevatia, R., Medioni, G.: ExpNet: landmark-free, deep, 3d facial expressions. In: IEEE Conference on Automatic Face and Gesture Recognition (2018)
11. Cosker, D., Krumhuber, E., Hilton, A.: Perception of linear and nonlinear motion properties using a FACS validated 3D facial model. In: ACM Applied Perception in Graphics and Vision (2010)
12. Cosker, D., Krumhuber, E., Hilton, A.: A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling. In: International Conference on Computer Vision (2011)
13. Cosker, D., Krumhuber, E., Hilton, A.: Perceived emotionality of linear and nonlinear AUs synthesised using a 3D dynamic morphable facial model. In: Proceedings of the Facial Analysis and Animation, FAA 2015, pp. 7:1–7:1. ACM (2015)
14. Dou, P., Shah, S.K., Kakadiaris, I.A.: End-to-end 3D face reconstruction with deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1503–1512, July 2017. https://doi.org/10.1109/CVPR.2017.164
15. Ferrari, C., Lisanti, G., Berretti, S., Del Bimbo, A.: A dictionary learning-based 3D morphable shape model. IEEE Trans. Multimedia **19**(12), 2666–2679 (2017). https://doi.org/10.1109/TMM.2017.2707341
16. Grupp, M., Kopp, P., Huber, P., Rätsch, M.: A 3D face modelling approach for pose-invariant face recognition in a human-robot environment. CoRR abs/1606.00474 (2016)
17. Hu, G., et al.: Face recognition using a unified 3D morphable model. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 73–89. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_5
18. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report 07–49. University of Massachusetts, Amherst, October 2007
19. Huang, Y., Khan, S.M.: DyadGAN: generating facial expressions in dyadic interactions. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2259–2266, July 2017. https://doi.org/10.1109/CVPRW.2017.280
20. Huber, P., Kopp, P., Rätsch, M., Christmas, W.J., Kittler, J.: 3D face tracking and texture fusion in the wild. CoRR abs/1605.06764 (2016). http://arxiv.org/abs/1605.06764
21. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)
22. Liarokapis, F., Debattista, K., Vourvopoulos, A., Petridis, P., Ene, A.: Comparing interaction techniques for serious games through brain-computer interfaces: a user perception evaluation study. Entertain. Comput. **5**(4), 391–399 (2014). https://doi.org/10.1016/j.entcom.2014.10.004. http://www.sciencedirect.com/science/article/pii/S1875952114000391

23. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: IEEE Conference on Computer Vision and Pattern Recognition-Workshops (2010)
24. Lüthi, M., Jud, C., Gerig, T., Vetter, T.: Gaussian process morphable models. CoRR abs/1603.07254 (2016). http://arxiv.org/abs/1603.07254
25. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. In: International Conference on Machine Learning (2009)
26. Masi, I., Ferrari, C., Del Bimbo, A., Medioni, G.: Pose independent face recognition by localizing local binary patterns via deformation components. In: International Conference on Pattern Recognition (2014)
27. Patel, A., Smith, W.A.P.: 3D morphable face models revisited. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
28. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3D face model for pose and illumination invariant face recognition. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 296–301 (2009)
29. Qiao, F., Yao, N., Jiao, Z., Li, Z., Chen, H., Wang, H.: Geometry-contrastive generative adversarial network for facial expression synthesis. CoRR abs/1802.01822 (2018). http://arxiv.org/abs/1802.01822
30. Ramanathan, S., Kassim, A., Venkatesh, Y.V., Wah, W.S.: Human facial expression recognition using a 3D morphable model. In: International Conference on Image Processing (2006)
31. Romdhani, S., Vetter, T.: Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In: IEEE Conference on Computer Vision and Pattern Recognition (2005)
32. Savran, A., et al.: Bosphorus database for 3D face analysis. In: Schouten, B., Juul, N.C., Drygajlo, A., Tistarelli, M. (eds.) BioID 2008. LNCS, vol. 5372, pp. 47–56. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89991-4_6
33. Tran, A.T., Hassner, T., Masi, I., Medioni, G.: Regressing robust and discriminative 3D morphable models with a very deep neural network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5163–5172, July 2017
34. Ujir, H., Spann, M.: Facial expression recognition using FAPs-based 3DMM. In: Tavares, J., Natal Jorge, R. (eds.) Topics in Medical Image Processing and Computational Vision. LNCS, pp. 33–47. Springer, Netherlands (2013). https://doi.org/10.1007/978-94-007-0726-9_
35. Yi, D., Lei, Z., Li, S.Z.: Towards pose robust face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
36. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.: A 3D facial expression database for facial behavior research. In: IEEE International Conference on Automatic Face and Gesture Recognition (2006)
37. Zakari, H.M., Ma, M., Simmons, D.: A review of serious games for children with autism spectrum disorders (ASD). In: Ma, M., Oliveira, M.F., Baalsrud Hauge, J. (eds.) SGDA 2014. LNCS, vol. 8778, pp. 93–106. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11623-5_9
38. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: a 3D solution. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)